Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Performance Comparison of Two Class Boosted Decision Tree snd Two Class Decision Forest Algorithms in Predicting Fake Job Postings

FHA. Shibly[1], Uzzal Sharma[2] and HMM. Naleer[3]

[1] Ph.D. Candidate, Assam Don Bosco University & Senior Lecturer, South Eastern University of Sri Lanka, shiblyfh@seu.ac.lk, Mob: 0094772301539

[2] Assistant Professor, Assam Don Bosco University, uzzal.sharma@dbuniversity.ac.in

[3] Senior Lecturer, South Eastern University of Sri Lanka, hmmnaleer@gmail.com

Abstract

During the Corona Virus Disease 2019 (COVID 19) period, online activities have become a necessary thing in everyone's life. However,in electronic recruitment, fake job postings have been started by scammers to get people's personal information and scam purposes. Many businesses prefer to post their vacancies electronically so that job applicants can access them quickly and timely. But this purpose may be one form of scam on the part of the fraud individuals because they give job applicants during terms of taking money from them or collecting their personal information for involving in cybercrimes. Fake job posting advertisements can be written against a reputable firm for breaching its reputation. The fraudulent post-detection work draws proper attention to obtaining an automated tool to identify fake jobs and report them to people to avoid applying for such situations. At present, many machine learning algorithms have been used to detect such fraudulent posts. But, the performance of such algorithms to be measured and compared to find a proper algorithm to incorporate in identifying fake things. In this research, the use of a proposed model with the help of Microsoft Azure Machine Learning Studio tested a comparison study on the performance of a two - class boosted decision tree and two - class decision forest algorithms. Researchers used F1 Score, Recall. Accuracy and precision to compare those two algorithms. Results showed that a two - class boosted decision tree is better for detecting fake job posts than the two - class forest decision algorithm. Thus, a two - class decision forest algorithm can be used to find and identify false or gossip messages, tweets, and social media publications.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Keywords:Two class decision forest, Fake job postings, machine learning, MS Azure and Two class boosted decision tree

Introduction

Nowadays, the internet and social media are becoming more critical things in people's life. People use these two technologies for most of their needs. All functions of organizations are being incorporated with technology, and computerized systems like enterprise applications, management information systems, and office automation applications are widely used in organizations. In Human Resource Manage, the Information System for Human Resources which is called as HRIS is one of among fully fledged electronic system for managing human resource activities and functions which is used by many organizations for making human resource functions efficiently.Moreover, recruitment and selection function of human resource management

Besides, the trend of applying for jobs online becomes more convenient for both employers and applicants. Most of the human resource agents and organizations enable the online application system for the recruitment and selection process. It has plenty of advantages. Applicants can apply with no time and easy to upload their curriculum vitae for further references. Also, they can useit anytime, anywhere. Employers also can filter the applications quickly and make shortlists within a short period. Therefore, electronic recruitment makes human resource functions speedy.

It provides a perfect opportunity for online scammers to take advantage of their desperation in these desperate times, when thousands and millions of people are on the lookout for a job. We see a daily increase in these fake job posts where posting seems quite normal, often these companies will also have a website, and they will have a recruitment process similar to other firms in the sector [1].

On the internet, there are a lot of job ads, even on the reputed work advertising pages, which never seem to be false. But the so-called recruiters start asking for the money and the bank details after the pick. A lot of the candidates slip into their trap and sometimes lose a lot of money and the current job. So, whether a job advertisement posted on the site is real or fake is better to identify [2].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Especially after the Corona Virus Disease 2019 (COVID 19), online activities are everywhere in everything. However, a dark side has become vulnerable to conduct electronic recruitment since fake job postings have been started by scammers to get people's personal information and scam purposes. Due to the unemployment rate, people are trying to apply all kinds of jobs which are suitable for them, and they don't think much about scamming.

It provides a perfect opportunity for online scammers to take advantage of their desperation in these desperate times when thousands and millions of people are on the lookout for a job. We see a daily rise in these fake job posts where posting seems quite reasonable, often these firms will also have a website, and they will have a recruitment process similar to other firms in the sector.

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF) [3].

Fake job postings available on recruitment websites until they are identified, sometimes taking more days to complete. Meanwhile, individuals innocently spend time and effort filling out job applications for fake sites with personal and confidential details. Moreover, if the contact details on the fake advertisement are those of a real business, it will have to deal with receiving vacancy applications that do not exist. Therefore, a proper mechanism should be identified and implemented automatically. Machine learning algorithms can help to find fake job postings, and it will be helpful for applicants to save their time and personal data without any troubles.

Research Problem

During past days, many businesses prefer to post their vacancies electronically so that job applicants can access them quickly and timely. But this purpose may be one form of scam on the part of the fraud individuals because they give job applicants during terms of taking money from them or collecting their personal information for involving in cybercrimes. Fake job posting advertisements can be written against a reputable firm for breaching its reputation. The fraudulent post-detection work draws proper attention to obtaining an automated tool to identify fake jobs and report them to people to avoid applying for such situations.

The machine learning approach is applied for this purpose, which uses multiple classification algorithms to recognize fake posts. A detection tool protects fake job posts from a more wide-rangingcollection of job ads in this scenario and alerts the user. Initially, the supervised learning

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

algorithm as the classification techniques are considered to address the problem of recognizing scams on the job posting. By considering training data, a classifier modelthe input variable to target classes. Classifiers discussed in the paper are briefly listed for recognition of fake work posts from the others. These classifiers - based predictions can be broadly classified into-Single Classifier - based Prediction and Ensemble Classifier - based Predictions [4].

Most of thealgorithms of machine learning and models of trained system can be used to recognize fake posts. The above two algorithms are used particularly in the classification of data. The Enhanced Decision Tree Two-Class builds a Master Learning model with an algorithm for boosted decision - making booms. An enhanced choice tree is an ensemble learning method that corrected a first tree loss, a second tree loss, a second tree loss etc. Predictions are based on each other on all the trees which predict. Furthermore, two - class forest decision algorithms are a group-based classification learning method. The overall principle is that we can findgood outcomes and a more common parameter instead of depending on one model by developing and incorporating several related models. Together models generally cover better decision making and better accuracy than individual decisions.There are very few research works have done to measure these two algorithms and a better model to be proved to detect fake job posts more accurately.

Related Works

The ensemble approach makes it possible to achieve a greater precision for the entire system by using a variety of machine learning algorithms together. Random Forest (RF) [5] uses a classification - based approach to learning and regression. Each classifier assimilates multiple tree-like classifiers applied to different samples, and each tree votes for the most appropriate class for entry. Boosting is a useful technique that incorporates multiple unstable learners into one learner to improve the precision of classification [4]. Expanded technology applies an algorithm for classifying the weighted versions of training data and selects the sequence of the majority voting classifiers. AdaBoost [6] is an excellent example of a technique of boosting, which produces better efficiency. Increasing algorithms means solving problems with spam filtration very effectively. Gradient boosting [7] is an additional boosting technique-a Classifier based on the decision tree principle. It also minimizes the loss of precision.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Fake news has malicious social media user accounts, repeating the impact of the room. The necessary investigation into counterfeit news is based on three points of view— how false news is written about how false news is distributed, how the user is connected to fake news. The news and social context characteristics have been extracted and a fake news learning model has been developed [8].

There were someresearchers used machine learning tools and techniques specially in the above two algorithms in detecting fake or rumors in social media. So, the researchers mainly focused on these algorithms and their performance in finding fraudulent things which can be converted to a systematic model for detecting fake job posts in social media.

A better algorithm machine learning method in which the second algorithm reviews and corrects the first tree failures, the third tree reviews and corrects the first tree and second tree errors, etc. Module Two - Class Enhanced Decision Tree provides an algorithm for enhanced decision trees as the basis for a machine learning model. A boosted Decision Tree is an ensemble machine learning model with a second tree that modifies the first tree errors, a third tree that adjusts the second tree, first and so on. Predictions are based on all trees, so that prediction can be made.

Another algorithm we can use in detecting fake things in social media is the decision forest. Decision forests are models of a fast, controlled ensemble. This module is the right choice if you want to predict an objective with up to two tests. We recommend using the Tune Model Hyperparameters framework for training and testing multiple models if you are confusion about the best results for a model decision tree. You will find the optimum solution by tuning iterations over a number of possibilities.

Therefore, this research is focused on the above algorithms in detecting fake job posts. By investigating these two algorithms, we can find a feasible and appropriate algorithm to find fraudulent job postings in social media.

Methodology

Proposed Approach

The next step was to divide the dataset into a training and testing dataset to train a model that can measure and compare the two algorithms.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Dataset

Before analyzing fake job postings, it is very important that we know what kind of post is wrong and what the extraction feature with this particular text that can categorize it as fake. Shivam has provided a data set containing 18 000 job descriptions, of whom approximately 800 are fake, in order to fulfill the prediction [8].

The data comprises of both meta-info and textual about employees. You can use the dataset to construct classification models to learn fake job descriptions.

Model

Researchers considered two prominent machine learning algorithms used to measure and compare fake postings: two class decision boosted tree (Algorithm 01) and two class decision forest algorithm (Algorithm 02). That model has been trained on training data by performing divisive data and train data in a machine learning lab in Microsoft Azure (MS Azure). On the basis of the accuracy, precision, recall & F1 results, the output of every algorithm is evaluated. Comparison is made of these two algorithms.

In this model, the selected dataset was undergone into preprocessing. It splits in the machine learning studio and undergoes to connect with algorithms. After that the model was trained, measure the score and evaluate the score. This model has been adopted with the guidelines given in MS Azure machine learning studio which is a readily available machine learning tool and to train and test a model. The value of accuracy determines the validity of the model. Since the dataset were already preprocessed and categorized, researchers didn’t practice any data cleansing and preprocessing tasks. But the model should be evaluated and the results should be compared. Data analysis proportion can be done with the machine learning studio which returns the necessary scores to measure and compare these two algorithms to obtain the objective of this research.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

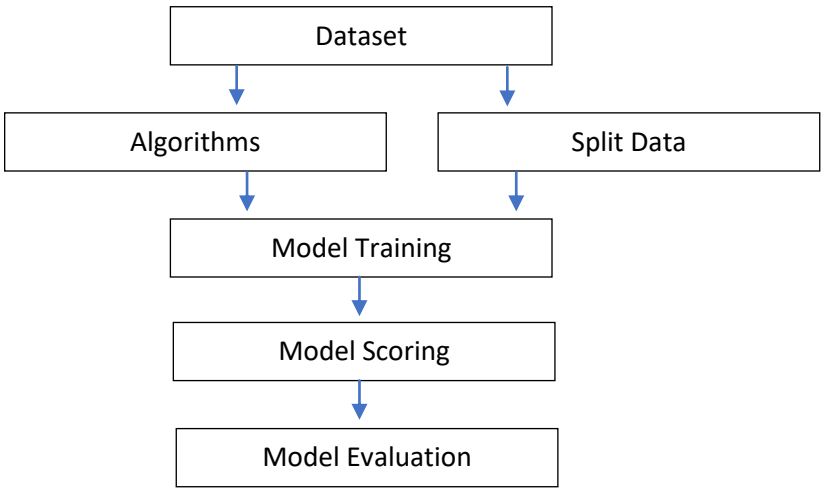

Fig 1. Proposed Approach

Results

The findings of the comparative analysis of algorithms are shown in the following,

Table 1:Comparison metrics

	Algorithm 01	Algorithm 02
Accuracy	0.938	0.954
Precision	0.720	1.000
Recall	0.750	0.020
F1 Score	0.735	0.039

According to Table 1, researchers have calculated the comparison metrics of two algorithms to compare their metrics.

The first is accuracy, and only a relationship of correctly predicted observe to total observation is the most intuitive measure of performance. One might think that our model is best if we are highly accurate. Precision is a significant measure, only if symmetrical datasets with nearly the same value of false positive and false negatives are available. So, to check the performance of our model, we will look at other parameters. Researchers have a model of 0.938 and 0.954, which is about 90 per cent exact, respectively for both algorithms for our model, which would be best.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Secondly, the precision is the percentage of correctly predicted positive findings of the total predicted positive findings. According to the above analysis, two-class decision forest algorithms have obtained 1.000 scores and it would be a perfect score equated to the two-class decision boosted tree.

Third, Recall is the proportion of properly predicted positive findings to all actual class observations. Two class boosted decision tree has a recall score of 0.750 based on the above analysis, and it is the best one then two - class decision forest algorithms which have reached 0.020.

Ultimately, the F1 Score for precision and recall is a weighted average. This ranking takes false positives and false negatives into account. It is not so simple to understand intuitively as accuracy, but F1 generally is more useful than precision, particularly if you have an unequal class distribution. Precision works best when the costs of incorrect positive and false negative are the same. If the value of false positive and false negative is somewhat different, consider both accuracy and warning in better terms. In our analysis the decision - making tree improved by F1 Two - class value is 0.735, and the two - class forest algorithms are 0.039.

Receiver Operating Characteristic curve(ROC)

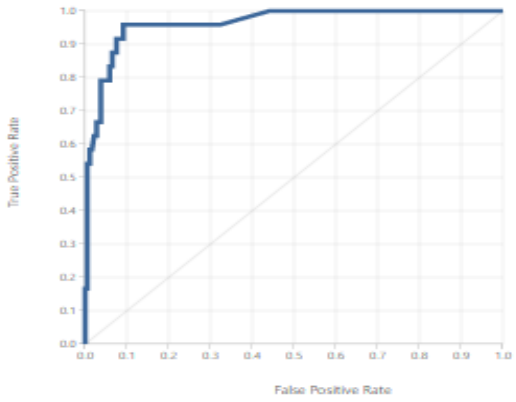

Fig 2 Algorithm 01

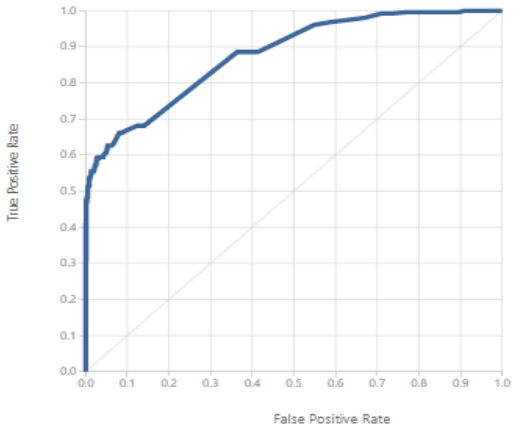

Fig 3Algorithm 02

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.953	0.000	1.000	0.000	0.953	1.000	0.000
(0.800,0.900]	0	0	0.000	0.953	0.000	1.000	0.000	0.953	1.000	0.000
(0.700,0.800]	0	0	0.000	0.953	0.000	1.000	0.000	0.953	1.000	0.000
(0.600,0.700]	2	0	0.000	0.953	0.016	1.000	0.008	0.953	1.000	0.000
(0.500,0.600]	3	0	0.001	0.954	0.039	1.000	0.020	0.954	1.000	0.000
(0.400,0.500]	8	0	0.002	0.955	0.097	1.000	0.051	0.955	1.000	0.000
(0.300,0.400]	18	0	0.006	0.958	0.218	1.000	0.122	0.958	1.000	0.000
(0.200,0.300]	40	2	0.014	0.966	0.434	0.973	0.280	0.965	1.000	0.000
(0.100,0.200]	71	102	0.046	0.960	0.568	0.577	0.559	0.978	0.980	0.010
(0.000,0.100]	112	5006	1.000	0.047	0.090	0.047	1.000	1.000	0.000	0.879

Fig 4 Algorithm 01

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	14	2	0.076	0.943	0.700	0.875	0.583	0.948	0.989	0.004
(0.800,0.900]	2	5	0.110	0.929	0.681	0.696	0.667	0.957	0.962	0.021
(0.700,0.800]	0	0	0.110	0.929	0.681	0.696	0.667	0.957	0.962	0.021
(0.600,0.700]	2	0	0.119	0.938	0.735	0.720	0.750	0.968	0.962	0.021
(0.500,0.600]	0	0	0.119	0.938	0.735	0.720	0.750	0.968	0.962	0.021
(0.400,0.500]	0	0	0.119	0.938	0.735	0.720	0.750	0.968	0.962	0.021
(0.300,0.400]	1	0	0.124	0.943	0.760	0.731	0.792	0.973	0.962	0.021
(0.200,0.300]	0	3	0.138	0.929	0.717	0.655	0.792	0.972	0.946	0.033
(0.100,0.200]	2	3	0.162	0.924	0.724	0.618	0.875	0.983	0.930	0.047
(0.000,0.100]	3	173	1.000	0.114	0.205	0.114	1.000	1.000	0.000	0.964

Fig 5 Algorithm 02

Fig. 2 and Fig. 3 shows the ROC curve of both algorithms. Classifiers which give curves closer to the top - left bendin the curve specifywell enough in performance according to the ROC curve. A two-class boosted decision tree has a better performance than the other algorithm, according to our analysis. Fig. 4 and Fig.5 have clearly explained all the relevant metrics and achievements of both algorithms in detecting fake job postings in social media. Based on the scores of Recall, F1 Score, Accuracy and Precision. of two algorithms, the presentation of a two-class boosted decision tree is better than the Two class decision forest algorithms.

Conclusion

In this research paper, researchers tried to measure the efficiency of two-class boosted decision tree and two-class decision forest algorithms in predicting fake job postings by using the proposed model. We dida performance measurement and comparison in finding fake posts of both algorithms on various sets of feature values and model hyperparameters. The findings and results showed that two-class boosted decision tree is healthier than the Two class decision forest algorithms in detecting fake job posts. Therefore, algorithm 01 can be used in finding and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

identifying fake or rumor posts, comments and publications in social media. It will return more reliable outputs than algorithm 02. In the future, more algorithms can be tested and compared to find more reliable parameters to detect fake things to control unnecessary burdens to social media users. The dataset and context also can be enlarged and enhanced to find more results in different approaches. Therefore, we can use and build new models by using two-class boosted decision tree to find or detect fraudulent posts specially job postings in social media or digital pages. This research will open up some more researches in the field of machine learning studio of Microsoft azure to train, test and compare performances of its algorithms. Since, most of the algorithms are readily available in the studio, researches and developers can build powerful models easily and quickly. There should be more researches to be done in the field of detecting fake posts in many aspects including fake news, fake profiles, fake messages and so on. This research definitely will help to find the performance and apply proper machine learning algorithms to create a safest online environment. The role of machine learning models and algorithms on detection of fake posts is immense and more researches will lead more new outcomes in the field of machine learning and fake detections.

References

[1]. Jain, S. (2020, April 24). Predicting Fake job postings-Part 1 (Data Cleaning & Exploratory Analysis). Retrieved from https://towardsdatascience.com/predicting-fake-job-postings-part-1-data-cleaning-exploratory-analysis-1bccc0f58110

[2]. Kumar, D. V. (2020, June 26). Classifying Fake and Real Job Advertisements using Machine Learning. Retrieved from https://analyticsindiamag.com/classifying-fake-and-real-job-advertisements-using-machine-learning/

[3]. Dutta, S., & Bandyopadhyay, S. K. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology, 68(4), 48-53. doi:10.14445/22315381/ijett-v68i4p209s

[4]. B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection," J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009

[5]. D. E. Walters, ―Bayes's Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2462 – 2472
Received 05 March 2021; Accepted 01 April 2021.

[6]. F. Murtagh, ―Multilayer perceptrons for classification and regression,‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[7]. L. Breiman, ―ST4_Method_Random_Forest Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[8]. B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, ―Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

[9]. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, ―Fake News Detection on Social Media,‖ ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[10]. Kaggle.com. 2020. [Real Or Fake] Fake Jobposting Prediction. [online] Available at: <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction> [Accessed 18 July 2020].